Update

Genome Analysis

# Estimating the size of the bacterial pan-genome

**Pascal Lapierre[1] and J. Peter Gogarten[2]**

[1] University of Connecticut Biotechnology Center, 91 North Eagleville Road, Storrs, CT 06269-3149, USA
[2] Department of Molecular and Cell Biology, University of Connecticut, 91 North Eagleville Road, Storrs, CT 06269-3125, USA

**The 'pan-genome' denotes the set of all genes present in the genomes of a group of organisms. Here, we extend the pan-genome concept to higher taxonomic units. Using 573 sequenced genomes, we estimate the size of the bacterial pan-genome based on the frequency of occurrences of genes among sampled genomes. Using gene- and genome-centered approaches, we characterize three distinct pools of gene families that comprise the bacterial pan-genome, each evolving under different evolutionary constraints. Our findings indicate that the pan-genome of the bacterial domain is of infinite size (the Bacteria as a whole have an open pan-genome) and that ~250 genes per genome belong to the extended bacterial core genome.**

## Genome plasticity and evolution

The availability of several hundred completely sequenced genomes has changed our views of genome evolution and uncovered extensive gene sharing between organisms. The view of stable genomes that function as unchanging information repositories has given way to a more dynamic view in which genomes frequently lose genes and incorporate foreign genetic materials [1,2]. The term 'pan-genome' or 'supragenome' denotes the set of all genes present in the genomes of members of a group of organisms, usually a species [3,4]. The pan-genome includes genes present in only one organism (known as ORFans), in the genomes of a few members of the group or in genes that are present in all genomes of the group (known as the core genome). Previously, Tettelin *et al.* [3] have shown that each individual strain of Group-B *Streptococci* (GBS) contains 13–61 unique genes and that, extrapolated to infinity, one would expect to find ~30 new genes for every additional GBS genome sequenced. Here, we apply this pan-genome concept to the bacterial branch of the tree of life, evaluating the dynamics of genome and gene family evolution and characterizing two modes of evolution: reuse with variation and *de novo* creation.

## From gene frequency to pan-genome

The approach developed by Tettelin *et al.* [3] to define the pan-genome consisted of tracking the number of unique genes among genomes in successive blast searches. This genome-oriented method is useful when a limited number of genomes are analyzed but computationally difficult when the number of genomes sampled is too large (total number of different sequential paths for $n$ genomes sampled is equal to $n!$). Because this method enables estimation of the frequency of occurrence of genes in genomes, the reverse also hold true. By using the frequency of occurrence of genes among genomes (i.e. in how many genomes do sampled genes have a homolog?), one can extrapolate back the sampling curve of the actual pan-genome of the group of organisms studied. This gene-oriented method has the advantage of being computationally less intensive and simultaneously providing a direct assessment of the gene frequencies among genomes, regardless of their genome of origin. Both approaches were initially compared using 293 completely sequenced genomes that were available at the time when this analysis was first conducted. The gene-oriented approach was later expanded to 573 bacterial genomes (for a list of all genomes sampled, see Table S1 in the supplementary material online) and yields very similar results (Table S2). We did not include archaeal genomes in our analyses because archaeal and bacterial homologs often are too divergent to establish homology through simple blast searches.

A total of 15 000 open reading frames (ORFs) were randomly selected from any of the 293 genomes (each ORF could only be selected once) and basic local alignment search tool (BLAST) searches were used to determine for each gene the number of genomes in which homologous sequences could be found (we required a bitscore >50 to classify a gene as present in the target genome and as a member of the same gene family). The total of 15 000 genes is sufficient to accurately reconstruct the sampling curve from the genome-centered approach. The resulting data were used to build a histogram in which each point represents the normalized number of genes ($An$) at the different frequencies (Fq) of occurrence in genomes (Figure 1a). The frequency distribution shows clustering of genes at both extremities of the histogram and most frequency categories contain approximately the same number of genes in the central part. The reconstruction of the sampling curve by adding up each individual component of the histogram, $f(x) = \sum[A_n{}^{*}e^{(Kn^{*}x)}]$, agrees with the data generated using the genome-centered approach, showing the equivalence of the two approaches (Figure 1b). From this histogram, three groups of ORFs are distinguished: (i) the extended core made up of ORFs on the right hand side of the diagram that occur in all or nearly all genomes; (ii) the accessory pool represented by ORFs on the left hand side of the diagram, comprising genes present in only one or a few genomes; and (iii) the remainder of the diagram comprised of proteins that are encoded in only a subset of the genomes. Here, we term these character genes because they define or can be used to define the character of groups of genomes. A decay function fitted to the reconstructed sampling curves was used to estimate

*Corresponding author:* Lapierre, P. (pascal.lapierre@uconn.edu)

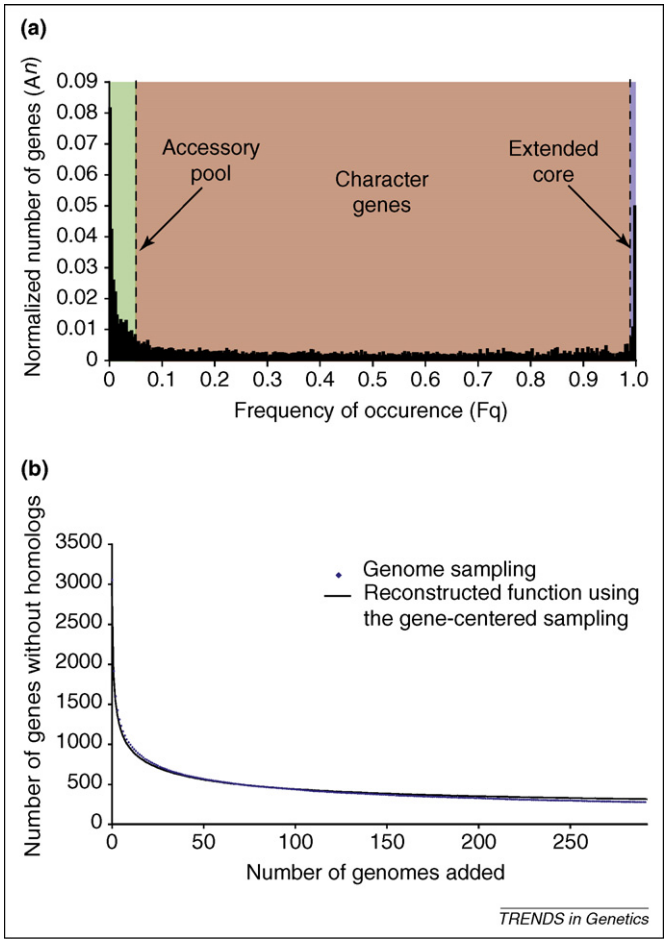


**Figure 1**. Frequency of occurrence of randomly selected genes in 293 bacterial genomes. **(a)** 15 000 genes were sampled to determine their frequencies of occurrence among genomes. Each bar corresponds to the normalized number of genes [*n* genes at Fq(*x*)/15000] having the indicated frequency (Fq) of occurrence (present in *n* other genomes/Total number of genomes −1). Genes without any homologs (Fq = 0) represent ORFans, whereas genes present in 292 other genomes (Fq = 1) represent strict core genes. Parts of the histogram that mainly contribute to the extended core, the character genes and the accessory pool are colored in blue, red and green, respectively (see Figure 2 for a definition of each of these categories). From the decay components (K) of the sampling functions for extended core and rare genes (see supplementary material online), the boundaries between the three pool of genes were determined by genes present in at least 99% of the genomes for the core set of genes and genes absent in at least 95% of the genomes for the accessory pool. **(b)** The frequency sampling can be used to reconstruct the sampling curve expected from the genome-centered approach. The sampling function reconstructed from the frequency histogram, $F(x) = \sum[A_n^* e^{(Kn^*x)}]$, $K = Ln(1 - Fq)$, agree with the data obtained with the sampling using the genome-centered approach. The slight difference between the genome-centered and the reconstructed sampling curves is caused by the probability of the sampling of individual gene. In the gene-centered approach, each gene, regardless of its genome of origin, has the same probability to be sampled, causing over representation of genes from larger genomes compared to the genome-centered approach. Because large genomes tend to harbor more duplicates and ORFans, it will cause the sampling curve to decay faster and to reach stability at slightly higher values.

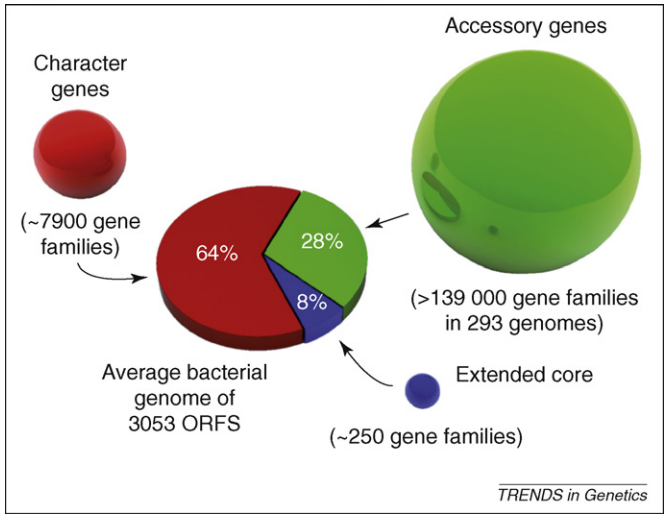


**Figure 2**. The bacterial pan-genome. Each gene found in the bacterial genome represents one of three pools: genes found in all but a few bacterial genomes comprise the extended core of essential genes (~250 gene families that encode proteins involved in translation, replication and energy homeostasis); the character genes (~7900 gene families) represent genes essential for colonization and survival in particular environmental niches (e.g. symbiosis and photosynthesis); and finally, the accessory genes, a pool of apparently infinite size, contains genes that can be used to distinguish strains and serotypes; the function of most genes in this category is unknown. At the genomic level, a typical bacterial genome is composed of ~8% of core genes, 64% of character genes and 28% of accessory genes. Although the character genes contain only 7900 gene families, they are the most abundant at the genomic level. Expanding the gene-centered approach to 573 bacterial genomes or sampling of 508 genomes, excluding highly reduced genomes, yields similar results (Table S2), except that the total number of families in the accessory pool is increased as expected for an open pan-genome.

the size of the three different groups of genes and to extrapolate the sampling curves to higher numbers of sampled genomes as additional genomes are sampled (see methods in the supplementary material online for more details).

## The extended core, character genes and accessory pool

The existence of a core set of genes present in all bacteria is testament to the conservative nature of evolution. Within several billions of years of bacterial evolution, no successful replacement of the core genes evolved in any of the lineages leading to the studied genomes. The core set of genes is under high selective pressure for a function that prevents drastic changes. The gene frequency approach presented here enables relaxing the core definition to include genes that are missing in only a small fraction of the genomes. This extended core of shared genes, which represent genes present in at least 99% of the sampled genomes (as determined by the fastest decay component of the sampling function), constitutes ~8% of the genes present in a typical bacterial genome (Figure 2). As pointed out by Koonin *et al.* [5], this set of core genes does not correspond to the minimal set of genes necessary for an organism to survive and thrive in nature. It is rather a backbone of essential components on which the rest of the genome is built.

Interestingly, although the character genes were found to be the main component of every bacterial genome (~64% of the total genes on average), this set of genes only contains ~7900 gene families. The rather small number of protein families found in the character pool is offset by the flexibility of these genes in their ability to adapt to new functions. Although similar on the sequence level, the character gene families demonstrate high diversity of substrate specificity. Instead of using a random process of creating new genes *de novo* to adjust to a situation, the limited number of character gene families indicates that the preferred mode of adaptation in bacteria consists of exploring new solutions from existing sequences via gene duplications, mutations and a mix and match assembly of modular proteins [6–9]. For example, the large gene family

of ABC transporters has diverse substrate specificities [6], which is caused by the substitutions in the periplasmic binding subunit [7,8]. Type I polyketide-synthases (PKSs) are large modular proteins that rely mainly on the combination of different protein sub-domains to achieve different functions [9]. Spreading in genomes through gene duplications and transfers, the seven characterized PKS domains assemble into multifunctional enzymes that synthesize many important secondary metabolites [9,10].

The creation of new protein folds might be reflected in the accessory pool. Many of these genes are ORFans (i.e. genes that do not have homologous sequences in other genomes). A closer analysis of the accessory genes found in *Escherichia coli* str. K12 substr. MG1655 reveals that, on average, these ORFs have a greater AT%, tend to be shorter and many are composed of insertion sequence (IS) elements and prophage sequences (see supplementary material online). We have found >139 000 rare gene families scattered throughout the bacterial genomes included in this study. The finding that the fitted exponential function approaches a plateau indicates an open pan-genome (i.e. the bacterial protein universe is of infinite size); a finding supported through extrapolation using a Kezdy-Swinbourne plot (Figure S3). This does not exclude the possibility that, with many more sampled genomes, the number of novel genes per additional genome might ultimately decline; however, our analyses and those presented in Ref. [11] do not provide any indication for such a decline and confirm earlier observations that many new protein families with few members remain to be discovered [12].

This set of accessory genes does not seem to be tightly bound to a particular organismal lineage. Their low level of conservation might indicate processes that can create new proteins [13]. These genes are frequently not subject to strong selective pressures [13,14] and they have high turnover rates in genomes [15]. Their likely association with bacteriophages and plasmids indicates that their evolution might transcend the organismal line of descent [16–18]. Regardless of their mode of insertion into bacterial genomes, the genes of the accessory pool seem to represent an ongoing gene creation process different from domain shuffling. Some of the genes in the accessory pool represent annotation artifacts resulting in ORFs that are not actually transcribed and translated. However, the number of falsely identified ORFs is usually estimated to be much smaller than the size of the accessory pool (1–4% [19] versus 28% of accessory genes per genome found in this study). In most instances, the process of gene creation might not lead to useful functions and the genes can be lost from the genomes. Occasionally, a new invention might arise from this cloud of genes and spread in and between populations as a result of the adaptive advantage provided, thereby moving the encoding gene to the pool of character genes.

Extending the pan-genome concept to higher taxonomic levels exacerbates the ambiguity in deciding if a gene should count as a new addition to the pan-genome or be considered as already present. This problem already exists for the pan-genome of a single species, especially in case of paralogs; however, for organisms belonging to different phyla, a protein with the same function might be so divergent that only the use of PSI blast or clustering might identify the homology [20–23]. Incorporating lineage-specific duplications and distinguishing them from ancient paralogs might be a useful extension to our gene and genome-centered classification schemes. A paralogous protein with an alignment score above the cut-off (a bitscore of 50), which is present in a target genome and which has lost the orthologous gene, would falsely cause the query gene to be considered to be present in the target genome. Under both approaches, the sampling of genes does not discriminate between orthologs and paralogs. However, because every gene is used as a query, paralogous genes present in the same genome are counted as separate families, resulting in ~8000 gene families in the character gene pool and ~250 gene families in the extended core. The choice of a simple blast hit cut-off to identify homologs might lead to falsely classifying a gene as different, just because it has diverged beyond recognition. This results in an underestimation of the number of genes in the extended core (two character gene families might be joined into a single family present in almost all bacteria), although within the bacterial domain divergence for core genes to score below a bitscore of 50 seems unlikely. Conversely, our simple approach to identify homologs probably overestimates the number of character genes: A small number of these character genes might have diverged below the chosen similarity cut-off and could be joined into a single family if conserved domains were used as classification. For example, many between-phyla comparisons of reaction center proteins from different photosynthetic bacteria score below the cut-off but, considering their similar fold and function, these should be considered as homologs. In other words, the surprisingly small number of families in the character gene pool would be even smaller.

## Concluding remarks

Since the completion of the first genomic sequence, we have come to appreciate the many forces acting on genome evolution [24]. The view of stable genomes functioning only as slowly changing repositories of genetic information gave way to a dynamic view whereby genomes function like collecting bins, continuously gaining and losing genes along the way. This constant rain of genetic material on genomes from a cloud of frequently transferred genes enhances the chance of survival of species by introducing variability in the population. We have identified three categories of gene that compose each genome: the extended core, the character genes and an accessory pool of genes. Proteins in these categories are evolving under different constraints and rules. Genes in the extended core are under high selective pressure and only minute changes at the sequence level are allowed. Although many instances of gene transfers have been documented, they mainly spread in populations through vertical inheritance. Gene duplication and domain shuffling are the preferred mode of evolution of the character genes. This set of genes enables organisms to quickly adapt to changing conditions and to exploit new niches. Of the three sets of genes, the character genes are the most likely to be transferred between organisms. The last category of genes consists

of genes with low levels of conservation, which are scattered at low frequencies throughout the bacterial domain. This accessory pool of genes might represent in part genes that had previous functions in genomes (now pseudogenes) but that are now stripped of selective pressure. These fast evolving genes, perhaps residing in phage genomes most of the time, explore sequence spaces and, occasionally, a new useful protein fold might arise from this pool and spread through populations.

How is protein space explored in biological evolution? *A priori*, two extreme points of view are possible: protein evolution is predominantly the result of selection and rearrangements of already existing proteins or protein evolution is an ongoing process in which new proteins evolve as the exploration of the protein landscape continues. Our results provide evidence for both processes operating in the bacterial world.

**Acknowledgements**

This work was supported by NASA Applied Information Systems Research (NNG04GP90G), NASA Exobiology (NAG5–12367) and NSF Microbial Genetics (MCB-0237197). We acknowledge the Bioinformatics services at the Biotech Center of the University of Connecticut for their computational support. We also thank the three anonymous reviewers for their constructive comments.

## Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2008.12.004.

## References

1 Snel, B. *et al.* (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* 12, 17–25
2 Koonin, E.V. and Wolf, Y.I. (2008) Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Res.* 36, 6688–6719
3 Tettelin, H. *et al.* (2005) Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proc. Natl. Acad. Sci. U. S. A.* 102, 13950–13955
4 Hiller, N.L. *et al.* (2007) Comparative genomic analyses of seventeen *Streptococcus pneumoniae* strains: insights into the pneumococcal supragenome. *J. Bacteriol.* 189, 8186–8195
5 Koonin, E.V. (2003) Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat. Rev. Microbiol.* 1, 127–136
6 Davidson, A.L. and Chen, J. (2004) ATP-binding cassette transporters in bacteria. *Annu. Rev. Biochem.* 73, 241–268
7 Nanavati, D.M. *et al.* (2005) Substrate specificities and expression patterns reflect the evolutionary divergence of maltose ABC transporters in *Thermotoga maritima*. *J. Bacteriol.* 187, 2002–2009
8 Fukami-Kobayashi, K. *et al.* (2003) Parallel evolution of ligand specificity between LacI/GalR family repressors and periplasmic sugar-binding proteins. *Mol. Biol. Evol.* 20, 267–277
9 Weissman, K.J. (2004) Polyketide biosynthesis: understanding and exploiting modularity. *Philos. Transact. A Math Phys. Eng. Sci.* 362, 2671–2690
10 Jenke-Kodama, H. *et al.* (2006) Natural biocombinatorics in the polyketide synthase genes of the actinobacterium *Streptomyces avermitilis*. *PLOS Comput. Biol.* 2, e132
11 Falkowski, P.G. *et al.* (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science* 320, 1034–1039
12 Yooseph, S. *et al.* (2007) The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families. *PLoS Biol.* 5, e16
13 Daubin, V. and Ochman, H. (2004) Start-up entities in the origin of new genes. *Curr. Opin. Genet. Dev.* 14, 616–619
14 Gogarten, J.P. and Townsend, J.P. (2005) Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* 3, 679–687
15 Lawrence, J.G. and Ochman, H. (1997) Amelioration of bacterial genomes: rates of change and exchange. *J. Mol. Evol.* 44, 383–397
16 Hendrix, R.W. *et al.* (2000) The origins and ongoing evolution of viruses. *Trends Microbiol.* 8, 504–508
17 Daubin, V. *et al.* (2003) The source of laterally transferred genes in bacterial genomes. *Genome Biol.* 4, R57
18 Hsiao, W.W. *et al.* (2005) Evidence of a large novel gene pool associated with prokaryotic genomic islands. *PLoS Genet.* 1, e62
19 Aggarwal, G. and Ramaswamy, R. (2002) *Ab initio* gene identification: prokaryote genome annotation with GeneScan and GLIMMER. *J. Biosci.* 27, 7–14
20 Tatusov, R.L. *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28
21 Enright, A.J. *et al.* (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30, 1575–1584
22 Zhang, Z. *et al.* (1998) Protein sequence similarity searches using patterns as seeds. *Nucleic Acids Res.* 26, 3986–3990
23 Harlow, T.J. *et al.* (2004) A hybrid clustering approach to recognition of protein families in 114 microbial genomes. *BMC Bioinformatics* 5, 45
24 Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science* 269, 496–512

0168-9525/$ – see front matter 
doi:10.1016/j.tig.2008.12.004 Available online 23 January 2009